Econometric Theory, **37**, 2021, 464–494.
doi:10.1017/S0266466620000213

IDENTIFYING MULTIPLE MARGINAL EFFECTS WITH A SINGLE INSTRUMENT*

CAROLINA CAETANO
University of Georgia
JUAN CARLOS ESCANCIANO
Universidad Carlos III de Madrid

This paper proposes a new strategy for the identification of the marginal effects of an endogenous multivalued variable (which can be continuous, or a vector) in a model with an Instrumental Variable (IV) of lower dimension, which may even be a single binary variable, and multiple controls. Despite the failure of the classical order condition, we show that identification may be achieved by exploiting heterogeneity of the "first stage" in the controls through a new rank condition that we term *covariance completeness*. The identification strategy justifies the use of interactions between instruments and controls as additional exogenous variables and can be straightforwardly implemented by parametric, semiparametric, and nonparametric two-stage least squares estimators, following the same generic algorithm. Monte Carlo simulations show that the estimators have excellent performance in moderate sample sizes. Finally, we apply our methods to the problem of estimating the effect of air quality on house prices, based on Chay and Greenstone (2005, *Journal of Political Economy* 113, 376–424). All methods are implemented in a companion Stata software package.

1. INTRODUCTION

Instrumental Variable (IV) methods are well established as one of the most useful approaches to identify causal effects in econometric models. Consider the nonparametric model

$$Y = g_0(X) + U, \tag{1}$$

where Y is the dependent variable, g_0 is an unknown measurable function of X, and U is an unobservable error term. The vector X is endogenous, in the sense

*We would like to thank Gregorio Caetano, Alon Bergman, Kenneth Chay, Matias Cattaneo, Pedro Sant'Anna, and and Robin Sickles for helpful discussions. We are thankful to the Co-Editor and three referees for suggestions, which have helped us to significantly improve the presentation. We would also like to thank seminar participants at many institutions as well as guests at our institutions for useful comments. This paper subsumes part of the working paper "Identifying Multiple Marginal Effects with a Single Binary Instrument or by Regression Discontinuity," dated May 25, 2015. Research funded to Juan Carlos Escanciano by the Spanish Grant PGC2018-096732-B-I00. Address correspondence to Carolina Caetano, Department of Economics, University of Georgia, 620 South Lumpkin Street Athens, GA 30602, USA; e-mail: carol.caetano@uga.edu.

© Cambridge University Press 2020

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

that $\mathbb{E}[U|X] \neq 0$ with positive probability. Depending on the nature of X and functional form assumptions on g_0, a traditional IV approach requires, among other things, that we observe an instrument Z that is sufficiently complex (see Newey and Powell, 2003). For instance, if X is continuous and we wish to identify g_0 nonparametrically, then Z must be continuous. If X is discrete with q points of support and we wish to identify all of its marginal effects, then Z needs to have at least q points of support. If X is a vector of continuous variables, then Z must have at least as many components as X. Thus, the traditional IV order condition imposes rather restrictive assumptions on the support of the instrument Z (relative to that of X), which often do not hold in applications. As a result, the marginal effects of X are not identified by traditional IV methods, for example, in common cases where the instrument is an experiment or a natural experiment.

To address this problem, a researcher may consider adding more variables to the list of instruments, say W, and run an IV regression with instruments (Z, W). However, if some component of W is correlated with U, then the IV regression with instruments (Z, W) would yield inconsistent estimates of the marginal effects of X on Y. In this paper, we propose an alternative identification strategy for marginal effects that allows for arbitrary correlation between W and U.

The starting point of our approach is the conditional exogeneity condition of Z given W, that is,

$$\mathbb{C}(U, Z|W) = 0 \text{ almost surely (a.s.)}, \tag{2}$$

where $\mathbb{C}(V_1, V_2|W)$ denotes the conditional covariance of V_1 and V_2 given W. Substituting U in the exogeneity condition yields the conditional covariance restriction

$$\mathbb{C}(Y, Z|W) = \mathbb{C}(g_0(X), Z|W) \text{ a.s.} \tag{3}$$

The main contribution of this paper is the use of covariance restrictions such as (3) for nonparametric identification under endogeneity. The paper introduces an identifying assumption, called *covariance completeness*, which provides conditions under which the right hand side of (3) can be uniquely solved (up to a constant) in g_0. For covariance completeness to hold W must be distinct from X (separability) and changes in Z must affect X differently according to W (differential effects). In contrast to the related completeness assumption of nonparametric IV (NPIV), our covariance completeness, which is extensively discussed below, imposes restrictions on the support of W relative to X (rather than on the support of Z relative to X, as in NPIV). As such, these necessary support conditions are more likely to hold in applications.

The following example illustrates and formalizes some of these ideas in a simple model where identification with a standard IV approach is not possible.

Example 1. (Linear model) Suppose that $X = (X_1, X_2)$ and g_0 is linear, so the model is

$$Y = \alpha_0 + \beta_{01} X_1 + \beta_{02} X_2 + U,$$

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

where $\mathbb{E}[U|X_1, X_2] \neq 0$. Here X_1 and X_2 can be two different variables, such as "education" and "completing a training program," or a relaxation of the linearity of a variable, for example, X_1 is "education," and $X_2 = \mathbf{1}(X_1 \geq 12)$ captures the "sheepskin effect of graduating from high school." Standard IV methods are unable to identify β_{01} and β_{02} with a single binary instrument Z, because the classical order condition fails.

For our approach, we use W, which may not be excluded from the structural equation. The exogeneity condition is that $\mathbb{C}(U, Z|W) = 0$. This condition allows, for example, that $U = \gamma_0 W + \varepsilon$, so W is potentially endogenous in the sense that $\gamma_0 \neq 0$, and the structural model becomes

$$Y = \alpha_0 + \beta_{01} X_1 + \beta_{02} X_2 + \gamma_0 W + \varepsilon, \tag{4}$$

where $\mathbb{C}(\varepsilon, Z|W) = 0$ a.s. yields

$$\mathbb{C}(Y, Z|W) = \beta_{01}\mathbb{C}(X_1, Z|W) + \beta_{02}\mathbb{C}(X_2, Z|W). \tag{5}$$

To identify β_{01} and β_{02}, we need to invert this equation. The condition that guarantees that we can invert it is what we call covariance completeness for the class $\mathcal{G} = \{g(X_1, X_2) = \beta_1 X_1 + \beta_2 X_2; \beta_1, \beta_2 \in \mathbb{R}\}$. In this example, covariance completeness holds if $\mathbb{C}(X_1, Z|W)$ and $\mathbb{C}(X_2, Z|W)$ are linearly independent. For example, if first stages are of the form

$$X_j = \alpha_{0j} + \alpha_{1j} Z + \alpha_{2j} W + \alpha_{3j} Z \cdot W + U_j, \qquad j = 1, 2, \tag{6}$$

with errors U_j satisfying $\mathbb{E}[U_j|Z, W] = 0$ a.s., then covariance completeness holds if

$$rank \begin{bmatrix} \alpha_{11} & \alpha_{31} \\ \alpha_{12} & \alpha_{32} \end{bmatrix} = 2, \tag{7}$$

$\mathbb{V}[Z|W] > 0$ a.s. and $\mathbb{V}(W) > 0$ (where $\mathbb{V}(\cdot)$ denotes variance). Our identification strategy exploits the heterogeneity in the "first stages" to separate the marginal effects of X_1 and X_2.

Our identification strategy has an intuitive interpretation. It is based on the observation that if the population is categorized according to W, the discrete variation on the distribution of X induced by Z may vary with W. This may allow us to recover a rich set of marginal effects even when Z is a binary variable, say $Z \in \{0, 1\}$, and X takes three or more values, and may even be continuous, or a vector. Thus, our results for binary IV open up the possibility of the identification of all the marginal effects of a variable X with rich support in cases where the instrument may be an experiment or a natural experiment.

Furthermore, we show that with almost no modifications our methodology can be extended to the nonseparable model

$$Y = m(X, U), \tag{8}$$

where $m(x, u)$ is a strictly monotone function in the scalar u, for each x in the support of the distribution of $X, \mathcal{S}_X$ say. Model (8) allows for unobserved

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

heterogeneous marginal effects, as in, for example, wage equations; see, for example, Card (2001). Example 7 in Section 4 provides an economic example where the production technology naturally implies the separability needed in our approach, while covariance completeness is plausible when the marginal cost of choosing X depends on interactions between W and Z (e.g., there is a differential effect of distance to college on education depending on parents' income).

We show that our identification strategy can be straightforwardly implemented with a simple two-stage least squares (TSLS) estimator that treats functions of W as exogenous controls and uses interaction terms between the same functions of W and a single Z as additional instruments. So, for example, in the bivariate linear model of Example 1, our estimator is a simple TSLS estimator with first stages (6) and a structural equation given by (4). We are able to extend the parametric TSLS estimator to semiparametric and nonparametric models by employing the general estimation results of Ai and Chen (2003) and Chen and Pouzo (2012, 2015). All the proposed estimators can be implemented as TSLS estimators with a common generic algorithm (see Section 3.3.1). A general purpose Stata package that implements our results is given in Caetano, Escanciano, and Bergman (2019).

The rest of the paper is organized as follows. First, we provide a review of the literature in the next section. We then introduce the ideas on the additive model, which is done in Section 3. There, we present the identification result in the binary IV case (Section 3.1), develop the concept of covariance completeness for the general case (Section 3.2), and show how to include covariates and propose robustness checks of separability (Section 3.2.1). We also show that the identification strategy can be implemented with a suitable TSLS estimator (Section 3.3) and provide a practical guide for implementation in Section 3.3.1. Section 4 extends the identification results to the nonseparable model. Section 5 reports some Monte Carlo experiments. Section 6 contains an empirical application of our method to the problem of estimating the effect of air quality on house prices, based on Chay and Greenstone (2005). Finally, we conclude in Section 7. Mathematical proofs of the main results are gathered in the Appendix.

2. LITERATURE REVIEW

Our identification results for separable models complement the classical NPIV approach in, for example, Newey and Powell (2003), Ai and Chen (2003), Darolles et al. (2011), Blundell, Chen, and Kristensen (2007), Horowitz (2011), and Chen and Christensen (2018). In the more general nonseparable case, our paper contributes to the literature of nonparametric identification of heterogeneous (observable and unobservable) marginal effects; see Matzkin (2013) for a survey of this literature. In particular, our results are complementary to alternative identification strategies for binary instruments and continuous endogenous variables in Chesher (2003), D'Haultfoeuille and Fevrier (2015), Torgovitsky (2015), D'Haultfoeuille, Holderlein, and Sasaki (2013), and Masten and Torgovitsky (2016); and for continuous instruments in Altonji and Matzkin (2005), Chernozhukov and Hansen (2005), and Florens et al. (2008), among others. Our identification strategy gives

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

rise to conditional moment restrictions that are estimated either by a standard TSLS estimator in the parametric case or by the general sieve Minimum Distance estimation methods of Ai and Chen (2003) and Chen and Pouzo (2012, 2015) in the semiparametric and nonparametric cases. In independent and contemporaneous work, Huang, Khalil, and Yildiz (2019) propose an identification strategy for situations with insufficient instruments. Their identification strategy is based on a control function approach and does not use covariates or covariance restrictions.

The main contribution of our paper relative to the aforementioned literature is the way we use covariates for identification. In particular, the papers mentioned above do not exploit the heterogeneity of the "first stage" conditional on possibly endogenous covariates nor discuss covariance completeness conditions, which are the main focus of this paper.

Covariance completeness is useful beyond the setting considered in this paper. In subsequent work to ours, Ben-Moshe, D'Haultfoeuille, and Lewbel (2017) and Kim and Song (2017) have used covariance completeness conditions to identify nonparametric models with mismeasured regressors and without instruments. The problem investigated by these authors is different from the one analyzed here. Nevertheless, these works illustrate the utility of covariance completeness in other applications of economic interest.

More broadly, our approach relates to traditional empirical methods in econometrics that treat controls as exogenous variables. If the controls turn out to be endogenous (see, e.g., experience in wage equations), then estimates of marginal effects of interest may be inconsistent; see, for example, Frolich (2008). We provide explicit sufficient conditions under which the consistency of the marginal effects of interest is not affected by the endogeneity of the controls both for our methods and also for traditional IV methods. Our semiparametric and nonparametric results relax the strong linearity assumptions of traditional IV approaches with control variables, see, for example, Chap. 12 of Stock and Watson (2011), but more importantly, suggests additional "instruments" that can be used to aid identification, even when these instruments include endogenous controls. Given the ubiquitous presence of controls in applications, this robustness is of practical relevance. Both our identification method and estimators have empirical advantages in that they can be adapted immediately to the idiosyncrasies of applied work, including incorporating functional form restrictions, very large number of controls and fixed effects all without the need to modify either the method or the estimator.

3. THE SEPARABLE CASE

Throughout this section, we assume that the observed random vector (Y, X, W, Z) follows the model

$$Y = g_0(X) + U \tag{9}$$

and satisfies the assumption:

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

Assumption 1. (i) $\mathbb{E}[Y^2] < \infty, \mathbb{E}[U^2] < \infty$ and $\mathbb{E}[Z^2] < \infty$; (ii) Z is exogenous in the sense that

$$\mathbb{C}(U, Z|W) = 0 \text{ a.s.} \tag{10}$$

Assumption 1(i) requires the existence of certain moments and Assumption 1(ii) repeats for convenience the maintained exogeneity assumption. An alternative exogeneity condition is

$$\mathbb{E}[U|W, Z] = \mathbb{E}[U|W] \text{ a.s.} \tag{11}$$

Our first result establishes the relation between (10) and (11). This discussion has important implications for our identification and estimation results. To that end, define the nuisance function

$$h_0(W) = \mathbb{E}[U|W],$$

and note that (10) can be equivalently written as the conditional moment restriction

$$\mathbb{E}[(Y - g_0(X) - h_0(W))\tilde{Z}|W] = 0 \text{ a.s.}, \tag{12}$$

where $\tilde{Z} = (1, Z)$. Alternatively, (11) gives rise to the moment restriction

$$\mathbb{E}[Y - g_0(X) - h_0(W)|W, Z] = 0 \text{ a.s.} \tag{13}$$

From (12) and (13), we obtain the following result (whose proof can be found in the Appendix).

PROPOSITION 3.1. *Let Assumption* 1(i) *hold. Then,*

(i) *The exogeneity condition* (11) *implies* (10).
(ii) *If Z is binary, then* (10) *also implies* (11).

Proposition 3.1 shows that the maintained conditional covariance exogeneity restriction is generally weaker than the exogeneity restriction (11), although both are equivalent when Z is binary. We use this equivalency below to investigate nonparametric identification of g_0 in the binary IV case. In our identification results, the functional form of the nuisance parameter $h_0(W) = \mathbb{E}[U|W]$ is unrestricted, and it is in this precise sense that we claim W to have arbitrary endogeneity.

Equivalently, we could write the exogeneity condition Assumption 1(ii) as the conditional moment restriction

$$\mathbb{E}[(Y - g_0(X))(Z - p_0(W))|W] = 0 \text{ a.s.},$$

where $p_0(W) = \mathbb{E}[Z|W]$ is the "propensity score." However, we prefer the representation in (12), as it has a direct link with our TSLS estimation strategy; see Section 3.3.

We also note that the traditional NPIV exogeneity assumption

$$\mathbb{E}[U|Z] = 0 \text{ a.s.} \tag{14}$$

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

is not maintained in our approach. If the researcher believes in this additional exogeneity condition, this conditional moment could be added to the moment restrictions in (12) to aid in the identification, improve the estimation method or test for overidentifying restrictions.

3.1. Identification with a Binary IV

Throughout this section, we assume that Z is binary, with support $\mathcal{S}_Z = \{0, 1\}$. From (13),

$$\mathbb{E}[Y|W, Z] = \mathbb{E}[g_0(X)|W, Z] + h_0(W).$$

Evaluating this expression at $Z = 1$ and $Z = 0$, and subtracting terms, we differentiate out the term $h_0(W)$ and obtain

$$\mathbb{E}[Y|W, Z = 1] - \mathbb{E}[Y|W, Z = 0] = \mathbb{E}[g_0(X)|W, Z = 1] - \mathbb{E}[g_0(X)|W, Z = 0] \text{ a.s.} \tag{15}$$

Identifying g_0 (up to location) from this equation depends on our ability to invert it. To better understand the conditions that guarantee the invertibility of equation (15) consider first the following example which clarifies how some of the requirements and ideas translate to an actual applied problem.

Example 2. (Discrete case) Consider the problem of estimating the marginal effect of the amount a woman smokes during pregnancy (average daily number of cigarettes) on the baby's weight at birth (see Almond and Currie, 2011 and Lumley et al., 2009 for discussions of the literature on this problem.) The following setup is entirely fictitious, but we believe that the association of our notation to a real problem can be helpful. Suppose that smoking can take three values $X \in \{0, 1, 3\}$. Later it will be immediate to see how the argument extends when X assumes more values. Suppose that women are randomly divided into two groups, indexed by Z, and let W be the mother's years of education. Table 1 shows an overview of the situation.

Column (I) represents the years of education, and at first we are considering only three possibilities: 6, 10, and 17. Columns (II) and (III) show the average amount smoked by the women in the control and treatment groups, respectively, for the given number of years of education. In this example, all groups reduced one cigarette on average because of the intervention. This is not a requirement of our method, we just want to show that identification can be achieved even when the "first stages" do not vary at all across the different values of W. Column (IV) shows the average difference (in grams) in the birth weight between the $Z = 1$ and the $Z = 0$ groups for that level of education ($\Delta_Y(W) = \mathbb{E}[Y|Z = 1, W] - \mathbb{E}[Y|Z = 0, W]$). Columns (V) to (VII) show the smoking distribution when $Z = 0$ ($\mathbb{P}_{Z,W}(x) = \mathbb{P}[X = x|Z, W]$). In this example, everyone in the $Z = 0$ group smokes three cigarettes, which is just a simplification for explanation purposes. Columns (VIII) to (X) show the corresponding fractions in the $Z = 1$ group. As we can see, each row is different. It means that Z affects each of the education groups differently. It is this variation

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

Table 1. Identification idea

(I) $\mathbf{W}$	(II) $\bar{\mathbf{X}}_{\mathbf{0},\mathbf{W}}$	(III) $\bar{\mathbf{X}}_{\mathbf{1},\mathbf{W}}$	(IV) $\Delta_{\mathbf{Y}}(\mathbf{W})$	(V) $\mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{0})$	(VI) $\mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{1})$	(VII) $\mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{3})$	(VIII) $\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{0})$	(IX) $\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{1})$	(X) $\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{3})$	(XI) Row #
6	3	2	10	0	0	1	0	1/2	1/2	(1)
10	3	2	22	0	0	1	1/5	1/5	3/5	(2)
17	3	2	30	0	0	1	1/3	0	2/3	(3)

in the distributions that is at the heart of our approach. It does not matter that all the average effects are the same, it would not even matter if there were no first stage effects at all. As we show next, our ability to identify the marginal effects comes from the fact that the instrument affected the distribution of X differently across the different W.

If Assumption 1 holds, from equation (15), we derive the following system of equations

$$10 = 0.5g_0(1) - 0.5g_0(3) \quad \textbf{(16)}$$

$$22 = 0.2g_0(0) + 0.2g_0(1) - 0.4g_0(3) \quad \textbf{(17)}$$

$$30 = 0.33g_0(0) - 0.33g_0(3). \quad \textbf{(18)}$$

Note that only two equations are linearly independent (since $0.4(16)+0.6(18) = (17)$). In fact, if we had used more values of the variable W, we could have more equations, but it would not change the fact that at most two equations would be linearly independent. This is caused by the fact that the coefficients of each of these equations always add up to zero, since they are the subtraction of probabilities, which themselves always add up to one.

Since we have two linearly independent equations, we cannot recover the values of $g_0(0)$, $g_0(1)$, and $g_0(3)$, but we can recover the value of any increment effect. It is straightforward to see in this example that, from equation (16), $g_0(3) - g_0(1) = -20$, from equation (18), $g_0(3) - g_0(0) = -90$, and combining both results, $g_0(1) - g_0(0) = -70$.

The discussion extends to the general discrete case as follows. Denote by $\mathcal{S}_X := \{x_1, \ldots, x_q\}$ and $\mathcal{S}_W := \{w_1, \ldots, w_l\}$ the supports of the distributions of X and W, respectively, with $q < \infty$ and $l < \infty$. Our identification strategy consists of inverting equation (15), which in this context can be written as a linear system $\Delta_Y = \mathbf{A}\mathbf{g}_0$, where $\Delta_Y := (\Delta_Y(w_1), \ldots, \Delta_Y(w_l))'$ (a' denotes the transpose of a) with $\Delta_Y(w) := \mathbb{E}[Y|Z=1, W=w] - \mathbb{E}[Y|Z=0, W=w], w \in \mathcal{S}_W$, the matrix $\mathbf{A}$ is given by $\mathbf{P}_1 - \mathbf{P}_0$, where $\mathbf{P}_z = (p_{zij})$ is the $l \times q$ matrix with entries $p_{zij} = \mathbb{P}[X = x_j | Z = z, W = w_i]$, $i = 1, \ldots, l$, $j = 1, \ldots, q$ and $z = 0, 1$, and $\mathbf{g}_0 := (g_0(x_1), \ldots, g_0(x_q))'$. Notice that since $\mathbf{P}_0$ and $\mathbf{P}_1$ are matrices of probabilities, $\mathbf{A}\iota = 0$, where ι denotes the $q \times 1$ vector of ones. Therefore, $\mathbf{A}$ is not full-rank, and thus g_0 is not identified from (15). However, in this context, we can identify linear functionals $c'\mathbf{g}_0$ with c in a space of dimension $rank(\mathbf{A})$. In particular, if $rank(\mathbf{A}) = q - 1$, then all

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

linear functionals $c'\mathbf{g}_0$ with $c'\iota = 0$ are identified. In this case, all increment effects $g_0(x_h) - g_0(x_j), h \neq j$, are identified. Of course, this is only possible if the order condition $l \geq q-1$ holds, so W needs to assume at least $q-1$ different values. The identification condition $rank(\mathbf{A}) = q-1$ is the key identification assumption in this paper and formalizes the idea of identification by exploiting heterogeneity of first-stages in covariates, as the elements of $\mathbf{A}$ are these effects, that is, $\mathbb{P}[X = x_j|Z = 1, W = w_i] - \mathbb{P}[X = x_j|Z = 0, W = w_i]$.

The discussion and intuition of Example 2 further extends to the general continuous case as follows. With some abuse of notation, we write equation (15) also as

$$\Delta_Y = Ag_0, \tag{19}$$

where now $Ag := \mathbb{E}[g(X)|W, Z = 1] - \mathbb{E}[g(X)|W, Z = 0]$ is a continuous (i.e., bounded) linear operator, $A : L_2(X) \to L_2(W)$, where henceforth, for a generic random vector $\zeta, L_2(\zeta)$ denotes the Hilbert space of square-integrable functions with respect to the distribution of ζ, with support $\mathcal{S}_\zeta$. We introduce our identification assumption as follows. Define $\mathcal{N}(A) = \{g \in L_2(X) : Ag = 0\}$, the null space of A. Our relevance condition requires that the null space of A is composed exclusively of the constant functions.

Assumption 2. (Relevance) $\mathcal{N}(A) = \{f \equiv c \in \mathbb{R}\}$.

Notice that the identification condition in Example 2 that $rank(A) = q-1$ is equivalent to Assumption 2 in the discrete support case, since $\dim(\mathcal{N}(A)) + rank(A) = q$ and $\mathcal{N}(A)$ of Assumption 2 has dimension one. The proof of the following identification result can be found in the Appendix.

THEOREM 3.2. *Under Assumptions 1 and 2, g_0 is identified in $L_2(X)$ up to location.*

Assumption 2 is the analogue in our setting of the L_2-completeness identification condition of NPIV, which is assumed throughout that literature; see for example, Newey and Powell (2003), Blundell et al. (2007), Andrews (2017), and D'Haultfoeuille (2011) for discussions on completeness. Sufficient and necessary conditions for Assumption 2 are given in the next section, as a special case of a more general result.

The intuitive differencing approach of this section does not have a natural extension to the case where Z takes more than two values. Moreover, as shown in Proposition 3.1, the exogeneity restriction (11) is more stringent than the simpler (10) for a general instrument Z. Based on these considerations, we introduce in the next section a nonparametric rank condition, called *covariance completeness*, which naturally generalizes our identification strategy to cases where Z is not binary, or where the structural equation is not separable, as we show later. Next, we define covariance completeness in detail and compare it to the classical completeness used in NPIV.

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

3.2. Identification in the General Case: Covariance Completeness

We introduce a general definition of covariance completeness and provide examples. To refine our identification result and provide sufficient and necessary conditions for identification, while allowing for the possibility of prior information on the parameter space for g, we introduce the following class of functions. Let $\mathcal{G}$ be a subset of $L_2(X)$, with $g_0 \in \mathcal{G}$, and whose elements satisfy the location normalization restriction $g(\bar{x}) = 0$, for a fixed $\bar{x} \in \mathcal{S}_X$, for all $g \in \mathcal{G}$. Define $\mathcal{G} - \{g_0\} = \{g - g_0 : g \in \mathcal{G}\}$.

Definition 3.1. *We say (X, Z) given W is $\mathcal{G}$-covariance complete if for each $g \in \mathcal{G} - \{g_0\}$,*

$$\mathbb{C}(g(X), Z|W) = 0 \text{ a.s.} \Longrightarrow g = 0 \text{ a.s.}$$

When $\mathcal{G}$ is unrestricted (except for the location normalization), we simply say (X, Z) given W is L_2-covariance complete or simply covariance complete.

The following result is proved in the Appendix.

THEOREM 3.3. *Let Assumption 1 hold. Then, g_0 is point-identified in $\mathcal{G}$ iff (X, Z) given W is $\mathcal{G}$-covariance complete.*

Theorem 3.3 generalizes Theorem 3.2 to the nonbinary IV case, as shown in the next result.

Corollary 3.1. *Assumption 2 is equivalent to covariance completeness if Z is binary.*

The following example illustrates how covariance completeness is satisfied for a familiar distribution.

Example 3. Suppose that (X, W) is jointly normal conditional on a binary IV Z, that is,

$$(X, W)|Z \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_Z \\ \rho_Z & 1 \end{pmatrix}\right).$$

Following Dunker et al. (2014), we can compute

$$\mathbb{E}[g(X)|W = w, Z = z] = (2\pi)^{-3/4} \exp\left(-\frac{w^2}{2}\right) \sum_{j=0}^{\infty} \mu^j(\rho_z) \mathbb{E}[g(X) p_j(X)] \frac{w^j}{\sqrt{j!}},$$

where p_j are the Hermite functions, $p_j(x) = (j! 2\pi)^{-1/2} \exp\left(-0.5x^2\right) H_j(x)$, with H_j the jth Hermite polynomial, and $\mu(\rho) = \rho/\sqrt{1-\rho^2}$. Therefore,

$$Ag(w) = (2\pi)^{-3/4} \exp\left(-\frac{w^2}{2}\right) \sum_{j=0}^{\infty} \left\{\mu^j(\rho_1) - \mu^j(\rho_0)\right\} \mathbb{E}[g(X) p_j(X)] \frac{w^j}{\sqrt{j!}}.$$

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

By the completeness of the Hermite polynomials, covariance completeness in this context translates into the intuitive condition $\rho_1 \neq \rho_0$, that is, variation in the instrument changes the correlation of X and W. We have been unable to extend this result to the nonzero conditional mean for $(X, W)|Z$, although we expect the result to go through in this case as well.

To compare covariance completeness with the classic concept of L_2-completeness, we define the latter formally; see Newey and Powell (2003), Blundell et al. (2007), Andrews (2017), and D'Haultfoeuille (2011) for further discussion on completeness.

Definition 3.2. *We say that the conditional distribution of R given S is $\mathcal{F}$-complete if for each $f \in \mathcal{F}$ the following holds*

$$\mathbb{E}[f(R)|S] = 0 \text{ a.s.} \Longrightarrow f = 0 \text{ a.s.}$$

When $\mathcal{F} = L_2(R)$, we simply say that the distribution of R given S is L_2-complete.

It is important to stress that the NPIV literature has used completeness of X given Z to identify marginal effects, which imposes stringent conditions on the support of Z relative to that of X (see the first paragraph in Section 1 for illustrations). The following result provides a sufficient condition for our covariance completeness in terms of a weighted completeness condition of X given W (rather than X given Z). Thus, our covariance completeness is fundamentally different from the completeness used in the NPIV literature, as they involve different conditioning variables. To the extent that controls have typically richer support than instruments, the support conditions required by our covariance completeness is more plausible than that of the standard NPIV approach.

Define $q(x, w) = \mathbb{E}[Z|X = x, W = w]$ and $k(x, w) = q(x, w) - p_0(w)$, where recall $p_0(w) = \mathbb{E}[Z|W = w]$ is the propensity score. Define also the class of measurable functions

$$\mathcal{F} = \{f(x, w) = g(x)k(x, w) : g \in \mathcal{G}\}. \tag{20}$$

The following result characterizes covariance completeness in terms of a weighted completeness condition (with weights given by the function $k(x, w)$).

PROPOSITION 3.4. *(X, Z) given W is $\mathcal{G}$-covariance complete iff the distribution of (X, W) given W is $\mathcal{F}$-complete, where $\mathcal{F}$ is the class defined in (20).*

This proposition is useful for several reasons. First, it shows that a necessary condition for covariance completeness is a nonzero weight function k, that is,

$$P(\mathbb{E}[Z|X, W] \neq \mathbb{E}[Z|W]) > 0. \tag{21}$$

This is a very intuitive necessary *relevance* condition: after controlling for W, X needs to be conditionally dependent with Z. Second, since $k(X, W)$ is uncorrelated with functions of W, this result shows that $\mathbb{E}[g_0(X)|W]$ is unidentified. This is

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

the separability condition. In the extreme case where $X = W, g_0$ is completely unidentified. Finally, since $\mathcal{F}$-completeness of R given S imposes necessary support conditions of S relative to R, Proposition 3.4 implies that covariance completeness imposes restrictions on the support of W relative to X. As mentioned before, this is in contrast to L_2-completeness of traditional NPIV, which imposes restrictions on the support of the instrument Z relative to X. In general, a necessary condition for covariance completeness is that both X and W have the same level of complexity (e.g., both are continuous), or more generally, the support of W is rich enough relative to that of X, as the following example, which subsumes previous ones, illustrates.

Example 4. (Examples 1 and 2, cont.) The intuition from Example 1 extends to the high-dimensional case. Assume $g_0(X) = \alpha_0 + \beta_0' \mathbf{X}$, where $\mathbf{X}$ is a possibly large dimensional vector of transformations of X of dimension d_x. This setting also covers the discrete example as a special case, by defining, for example, $\alpha_0 = g_0(x_1), \beta_{0j} = g_0(x_j) - g_0(x_1)$, and $\mathbf{X} = (1(X = x_2), \ldots, 1(X = x_q))$, so $d_x = q - 1$. Taking as the location normalization $\alpha_0 = 0$, covariance completeness holds if

$$\mathbb{C}(\mathbf{X}, Z|W)' \beta_0 = 0 \Longrightarrow \beta_0 = 0. \tag{22}$$

The precise necessary and sufficient condition for this is

$$\mathbb{E}[\mathbb{C}(\mathbf{X}, Z|W)\, \mathbb{C}(\mathbf{X}, Z|W)'] \text{ is positive definite.} \tag{23}$$

Thus, covariance completeness requires that first-stages are heterogenous in covariates. Specifically, in this example is needed that $\mathbb{C}(\mathbf{X}, Z|W)$ is not perfectly multicolinear, so that (23) holds. Of course, a necessary condition for covariance completeness is that $\mathbb{C}(\mathbf{X}, Z|W)$ varies with W (with at least d_x distinct values). This condition can be tested. For example, suppose $Z = \gamma + \delta' \mathbf{X} + V$, where $\mathbb{E}[V|W, X] = \mathbb{E}[V|W]$ a.s. Then, $\mathbb{C}(\mathbf{X}, Z|W) = \delta' \mathbb{V}(\mathbf{X}|W)$ and the necessary conditions are that $\delta \neq 0$ and conditional heteroskedasticity, so that $\mathbb{V}(\mathbf{X}|W)$ is not constant. Both are intuitive and can be tested by traditional methods. The condition $\delta \neq 0$ is the relevance condition mentioned earlier ($\mathbb{E}[Z|X, W] \neq \mathbb{E}[Z|W]$). There is no need to check these conditions separately. Indeed, we recommend checking variability of $\mathbb{C}(\mathbf{X}, Z|W)$ in W by testing significance of interaction coefficients in our first-stage regressions. Suppose that

$$\mathbf{X} = \alpha_0 + \alpha_1 Z + \alpha_2 \mathbf{W} + \alpha_3 Z \cdot \mathbf{W} + U_x,$$

where $\mathbf{W}$ is a possibly large dimensional vector of transformations of W of dimension d_w, α_0 and α_1 are d_x-dimensional vectors, α_2 and α_3 are $d_x \times d_w$ matrices of first-stage coefficients, and $\mathbb{C}(U_x, Z|W) = 0$. Denoting $\tilde{\boldsymbol{\alpha}} = (\alpha_1, \alpha_3)$ and $\tilde{\mathbf{W}} = (1, \mathbf{W})$, since

$$\mathbb{E}[\mathbb{C}(\mathbf{X}, Z|W)\, \mathbb{C}(\mathbf{X}, Z|W)'] = \tilde{\alpha} \mathbb{E}[\tilde{\mathbf{W}} \tilde{\mathbf{W}}' \mathbb{V}[Z|W]] \tilde{\alpha}',$$

then covariance completeness holds if

$$rank\left[\tilde{\alpha}\right] = d_x, \tag{24}$$

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

and

$$\mathbb{E}[\tilde{\mathbf{W}}\tilde{\mathbf{W}}'\mathbb{V}[Z|W]] \text{ is positive definite.} \quad \textbf{(25)}$$

The rank condition (24) extends (7) to the high-dimensional case and requires the order condition $d_w \geq d_x - 1$. The rank condition (25) is mild and depends on the joint distribution of controls and instruments.

In the discrete case for W, one can take $\mathbf{W} = (1(W = w_2), \ldots, 1(W = w_l))$, so $d_w = l - 1$, and the previously derived rank condition $rank(\mathbf{A}) = q - 1 = d_x$ is equivalent to (24), provided $\mathbb{V}[Z|W] > 0$ a.s. (so (25) is satisfied with a diagonal matrix).

3.2.1. *Relaxing and Testing Structural Separability.* A necessary condition for covariance completeness is that W must be separable from X. It is important to notice that one can have nonseparable covariates in the model. Suppose that Q is a vector of covariates, then the extended model

$$Y = g_0(X, Q) + U,$$

with the multivariate exogeneity condition

$$\mathbb{C}(U, Z|W) = 0 \text{ and } \mathbb{C}(U, Q|W) = 0$$

can be treated identically to the univariate case, simply replacing Z there by (Z, Q). The restrictions on the support of W relative to X remain unchanged. We discuss further this extension in Section 3.3.

Note that if we assume a semiparametric structure for g, for example, if $g(X, Q) = \beta(Q)'X$, with $\beta(\cdot)$ fully nonparametric, then we can even drop the exogeneity requirement for Q, as we show in the following example. We keep our notation for potentially endogenous controls and denote Q as W_2, where $W = (W_1, W_2)$.

Example 5. (Linear model with heterogeneous effects in W_2) Consider the varying coefficient model

$$Y = \alpha_0(W_2) + \beta_0(W_2)'X + U, \quad \mathbb{C}(U, Z|W) = 0 \text{ a.s.}$$

To simplify notation, for a generic random vector ξ denote

$$C_\xi(w_1, w_2) = \mathbb{C}(\xi, Z|W_1 = w_1, W_2 = w_2).$$

With this notation at hand, we derive from the conditional exogeneity the condition

$$C_Y(w_1, w_2) = C_X(w_1, w_2)'\beta_0(w_2).$$

Fixing w_2, evaluating w_1 at W_1, multiplying both sides by $C_X(W_1, w_2)$ and taking expectations, then identification of $\beta_0(w_2)$ holds if

$$\mathbb{E}[C_X(W_1, w_2)C_X'(W_1, w_2)] \text{ is positive definite.} \quad \textbf{(26)}$$

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

In this example, covariance completeness reduces to (26). Under this condition, we can estimate $\beta_0(w_2)$ nonparametrically from local least-squares regressions. Here, identification requires that W_1 is not empty and must assume at least as many different values as the dimension of X.

Alternatively, we could specify $\beta_0(W_2)$ as a linear function of W_2, say $\beta_0(W_2) = \beta_{01} + \beta_{02}' W_2$, which results in a linear-in-parameters model with endogenous variables X and endogenous interactions between X and W_2. In this setting, we could even have W_1 empty, so $W = W_2$. Consider, for example, the case where X is univariate. The model with interactions in the structural equation is

$$Y = \alpha_0 + \beta_{01} X + \beta_{02}' WX + U, \quad \mathbb{C}(U, Z|W) = 0 \text{ a.s.},$$

and our covariance completeness condition becomes

$$\mathbb{E}\left[\begin{pmatrix} C_X(W) \\ WC_X(W) \end{pmatrix} \begin{pmatrix} C_X(W) & WC_X(W) \end{pmatrix}\right] \text{ is positive definite.}$$

This model is just identified under the rank condition.

In the previous example, note that since the parameter of the interaction term β_{02} is identified, our method allows us to test for separability in the marginal effects, that is, $H_0 : \beta_{02} = 0$, for example, with a simple Wald test. We implement tests for separability following this approach in our application of Section 6. Since structural separability is an important component of our approach, we recommend performing tests of separability when applying our procedure. If some of the coefficients in β_{02} are zero, the model becomes over-identified under the rank condition, which opens up the possibility of testing for the over-identifying restrictions and moreover, makes estimation more robust. An example in the next section illustrates these points in the context of our estimator.

3.3. Estimation

The exogeneity condition $\mathbb{C}(U, Z|W) = 0$ a.s. leads to the following conditional moment representation of our model, which we repeat here for convenience (cf. 12):

$$Y = g_0(X) + h_0(W) + \varepsilon,$$

where

$$\mathbb{E}[\varepsilon \tilde{Z}'|W] = 0 \text{ a.s.}, \tilde{Z} = (1, Z). \tag{27}$$

These conditional moment restrictions can be used for estimation. If $g_0(X)$ and $h_0(W)$ are specified parametrically, then a Generalized Method of Moments estimator can be easily constructed. A simple case arises when $g_0(X) = c_0 + \beta_0' \mathbf{X}, h_0(W) = \tau_0 + \gamma_0' \mathbf{W}$, and where $\mathbf{X}$ and $\mathbf{W}$ are d_x and d_w-dimensional vectors of transformations of X and W, respectively, excluding an intercept. These specifications lead to a structural equation that is linear in parameters, that is,

$$Y = \alpha_0 + \beta_0' \mathbf{X} + \gamma_0' \mathbf{W} + \varepsilon, \tag{28}$$

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

where $\alpha_0 = c_0 + \tau_0$. Note that the intercept c_0 is generally not identified, but this is of no concern, since we are interested in the parameter β_0. In this section, we use the location normalization $c_0 = 0$ for identification of g_0.

The new insight of this paper is that the moment restriction (27) implies that new instruments can be constructed from interacting Z with transformations of W, and this should be compared with textbook treatments of the topic which suggest just using Z as instruments and W as controls, see, for example, Chap. 12 in Stock and Watson (2011). The simplest way to incorporate heterogeneity in first-stages X is through

$$\mathbf{X} = \alpha_{0X} + \alpha_{1X} Z + \alpha_{2X} \mathbf{W} + \alpha_{3X} \mathbf{W} Z + U_x,$$

where $\mathbb{E}[U_x|Z, W] = 0$ a.s. and Z is a single instrument. The coefficients α_{0X} and α_{1X} are $d_x \times 1$, while α_{2X} and α_{3X} are matrices $d_x \times d_w$. Of course, other first-stages could be used, but since $\mathbf{W}$ is potentially high dimensional, we believe this to be a sufficiently rich specification in practice.

With these specifications, our identification strategy can be easily implemented as a TSLS estimation, since as suggested by the moment condition (27)

$$\mathbb{E}[\varepsilon(\tilde{Z} \otimes q(W))] = 0,$$

for any measurable function q such as the moment is well-defined and where $\otimes$ denotes the Kronecker product. Using $q(W) = \tilde{\mathbf{W}} = (1, \mathbf{W})$ leads to $2(1 + d_w)$ unconditional moments with "exogenous variables" $S = \tilde{Z} \otimes \tilde{\mathbf{W}}$. Given that we have $1 + d_x + d_w$ parameters in the structural equation, the necessary order condition for a TSLS identification strategy is $d_w \geq d_x - 1$.

Specifically, given an independent and identically distributed (i.i.d.) sample $\{(Y_i, X_i, W_i, Z_i)\}_{i=1}^n$ of (Y, X, W, Z), we propose to estimate β_0 with the coefficient of the $\mathbf{X}_i$ on a TSLS regression of Y_i onto $\mathbf{X}_i$ and $\mathbf{W}_i$, using Z_i and $Z_i \mathbf{W}_i$ as instruments for $\mathbf{X}_i$, and treating $\mathbf{W}_i$ as if they were exogenous controls. This estimator, which we denote by $\widehat{\beta}$, can be implemented with off-the-shelf econometric software. For a practical guide about its implementation, see Section 3.3.1.

We could also entertain semiparametric and nonparametric specifications for g_0 and h_0, and use the general estimation results of Ai and Chen (2003) and Chen and Pouzo (2012, 2015) to estimate marginal effects or other functionals of g_0. For example, a semiparametric specification is obtained if $g_0(X) = \beta_0' \mathbf{X}$ and $h_0(W)$ is left unspecified. To describe the semiparametric estimator, let $\mathcal{H} \subset L_2(W)$ denote the parameter space for h_0, and let $\mathcal{H}_n \subset \mathcal{H}$ be a linear sieve parameter space

$$\mathcal{H}_n = \left\{ h \in L_2(W) : h(w) = \sum_{j=1}^{J_n} a_j q_j(w) \right\}, \tag{29}$$

where $q^{J_n}(w) = (q_1(w), \ldots, q_{J_n}(w))'$ is a vector containing a linear sieve basis, with $J_n \to \infty$ as $n \to \infty$, with the property that linear combinations of the $q_j's$ can approximate any square integrable function of W well. Assume the first element of the basis is a constant, that is, $q_1(w) \equiv q_1$. The sieve minimum distance estimator

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

of Ai and Chen (2003) for β_0 with an identity weighting is just the TSLS estimator discussed above with $\mathbf{W} = (q_2(w), \ldots, q_{J_n}(w))'$ and $d_w = d_{w,n} = J_n - 1$. Thus, under the conditions specified in Ai and Chen (2003), the semiparametric TSLS estimator for β_0 is root-n consistent and asymptotically normal. This estimator is not efficient, however, but an efficient estimator can be constructed by a suitable choice of the weighting matrix, as discussed in Ai and Chen (2003). We refer to this paper for details on the efficient estimator based on our conditional moments.

Likewise, nonparametric specifications of $g_0(X)$ could be also entertained and a nonparametric sieve approach will correspond to using $\mathbf{X} = (p_2(x), \ldots, p_{K_n}(x))'$ above, with $d_x = d_{x,n} = K_n - 1$, where $p^{K_n}(x) = (p_1(x), \ldots, p_{K_n}(x))'$ is a vector containing a linear sieve basis for $\mathcal{G} \subset L_2(X)$, the first element $p_1(x)$ being a constant term which is excluded from $\mathbf{X}$ as a location normalization, and $K_n \to \infty$ as $n \to \infty$. Here, the results in Ai and Chen (2003) and Chen and Pouzo (2012) deliver rates of convergence for the resulting TSLS estimator for g_0, and the results of Chen and Pouzo (2015) deliver inference for functionals of g_0 allowing for a growing dimension $K_n \to \infty$ as $n \to \infty$. In all these cases, the implementation as a TSLS estimator remains the same, with the standard TSLS necessary order condition $d_w \geq d_x - 1$, (i.e., $J_n \geq K_n - 1$).

The following example illustrates that our setting can also accommodate interactions between endogenous variables and controls, leading to simple tests for separability.

Example 6. (Example 1, cont.) Consider the linear model of Example 1 in Section 1, with two endogenous variables $X = (X_1, X_2)$, a d_w-dimensional vector of controls W and one IV, Z. Our estimator is the TSLS estimator applied to the structural model

$$Y = \alpha_0 + \beta_{01} X_1 + \beta_{02} X_2 + \gamma_0' W + \varepsilon, \ \mathbb{C}(\varepsilon, Z|W) = 0$$

with first stages

$$X_j = \alpha_{0j} + \alpha_{1j} Z + \alpha_{2j} W + \alpha_{3j}' Z \cdot W + U_j, \qquad j = 1, 2.$$

Although this model is under-identified with classical IV, it is identified with our approach under a standard TSLS rank condition. If $d_w > 1$, then an applied researcher can use our method to test for over-identifying restrictions by standard methods in TSLS, after estimating the model. Furthermore, if the researcher is concerned with the structural separability of W and the endogenous variables, as required in our method, we recommend, as robustness checks, running our estimator on the structural model with interactions

$$Y = \alpha_0 + \beta_{01} X_1 + \beta_{02} X_2 + \beta_{03} W_1 X_1 + \gamma_0' W + \varepsilon,$$

where W_1 is a univariate component of W and the same first stages as above. The estimation of this extended model follows our generic implementation with $\mathbf{X} = (X_1, X_2, W_1 X_1)$ and $\mathbf{W} = W$. The order condition $d_w \geq d_x - 1$ holds if $d_w > 1$, because $d_x = 3$. A standard t-test for $\beta_{03} = 0$ based on our TSLS estimator

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

provides empirical evidence of the separability (or lack thereof) used in the first specification. We can apply a similar argument with interactions with X_2 and components of W. The next section provides specific guidance on how to apply and interpret our methods.

Finally, we note that the extension of our results to the multivariate instrument case is straightforward. Collect in a d_z-dimensional vector $\mathbf{Z}$ all the instruments. Then, apply our TSLS estimator above where the vector of exogeneous variables in the first stages is $\tilde{\mathbf{Z}} \otimes \tilde{\mathbf{W}}$, where $\tilde{\mathbf{Z}} = (1, \mathbf{Z})$. The necessary order condition is now $d_z(d_w + 1) \geq d_x$. The elements in $\mathbf{Z}$ may be transformations of a basic instrument Z, corresponding to a strengthening of the exogeneity condition.

3.3.1. *Guide to Practice.* This section provides a step-by-step implementation of our TSLS estimator. We first give the steps and then discuss each of them separately. A general purpose Stata package that implements these steps is given in Caetano et al. (2019). We propose the following basic algorithm for implementation:

Step 1. Choose a small set of the endogenous variables of interest $\mathbf{X}$ of dimension d_x.

Step 2. Select the control variables in $\mathbf{W}$ with dimension $d_w \geq d_x - 1$. Set $\tilde{\mathbf{W}} = (1, \mathbf{W})$.

Step 3. Run a TSLS regression of Y onto $\mathbf{X}$ and $\tilde{\mathbf{W}}$, using $\tilde{\mathbf{Z}} \otimes \tilde{\mathbf{W}}$ as exogenous variables for $\mathbf{X}$.

Step 4. Apply standard TSLS inference, including standard error calculations and practical recommendations regarding weak instruments.

Step 5. Run diagnostic and sensitivity tests, as discussed below.

The first step in our procedure is the choice of the elements in $\mathbf{X}$. These elements could be distinct basic endogeneous variables, such as "education" and "completing a training program" as in our introductory example, or different transformations of a basic endogeneous variable, as in the discrete and nonparametric cases. In any case, we recommend choosing a small set first, for example, $d_x \leq 3$.

For Step 2, candidates for elements of $\mathbf{W}$ are controls, distinct from the variables in $\mathbf{X}$, that are thought as helpful for the plausibility of the conditional exogeneity assumption and lead to significant heterogenous effects in the first stages, so the rank condition (24) is satisfied. Note that not all the interaction effects need to be significant for (24) to hold. Note also that although one could estimate the matrices involved in the rank conditions (24) and (25) and check their ranks, packaged software will check these conditions implicitly when running TSLS. If completeness does not hold, the program will report an ill-conditioned inverse in the TSLS estimation.

Steps 3 and 4 are standard. Step 5 consists of robustness checks addressing concerns of misspecification of the structural equation, such as functional forms

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

misspecifications for g_0 and h_0 and checks of separability, as well as sensitivity to the specification of controls. We recommend first fixing a set of basic elements in $\mathbf{X}$. Then, increase the number of terms of $\mathbf{W}$, which corresponds to the semiparametric estimator of Ai and Chen (2003) where g_0 is linear and h_0 is unspecified. To the best of our knowledge, there are no theoretical results for data-driven choices of $d_w = J_n - 1$ and $d_x = K_n - 1$ in the context of our semiparametric and nonparametric specifications. Horowitz (2014) proposed a theoretically justified data-driven method for choosing the number of terms in a sieve NPIV context. Theoretical developments along this line are beyond the scope of this paper. In practice, we recommend (generalized) cross-validation for choosing J_n and K_n, see, for example, Wahba (1990), and then running some sensitivity checks by varying J_n and K_n subject to the constraint $J_n \geq K_n - 1$.

Another type of misspecification comes from the lack of separability between X and W. This can be also addressed by increasing the number of elements in $\mathbf{X}$, adding interaction terms such as $W_1 X_1$ to the original formulation of $\mathbf{X}$, as shown in the previous example, and testing for significance in the interaction coefficients. The maximum number of interactions that can be added is the number of original overidentifying restrictions $d_w - d_x + 1$.

We also note that not all controls need to be interacted with the instrument. Suppose that $\mathbf{Q}$ is a vector of exogenous controls as discussed in Section 3.2.1, Step 3 can also be implemented as a TSLS regression of Y onto $\mathbf{X}, \tilde{\mathbf{W}}$, and $\mathbf{Q}$, using $\tilde{\mathbf{Z}} \otimes \tilde{\mathbf{W}}$ as exogenous variables for $\mathbf{X}$.

4. NONSEPARABLE CASE

This section extends our previous identification strategy to the nonseparable model

$$Y = m_0(X, U), \tag{30}$$

where $m_0(x, u)$ is a strictly monotone function in a scalar u, for each x in $\mathcal{S}_X$. The following example motivates our structural separability and monotonicity assumptions in an economic setting.

Example 7. Consider the following example of production with an intermediate unobserved input. An economic agent produces some final outcome Y, such as firm revenue or individual lifetime earnings, following the steps:

- Step 1: W and ε are used to produce $U = q(W, \varepsilon)$.
- Step 2: Given W and ε_x, the agent chooses X

$$X = s(W, Z, \varepsilon_x) \equiv \arg\max_{x^*} \left\{ \mathbb{E}[m_0(x^*, q(W, \varepsilon)) | W, \varepsilon_x] - c(x^*, W, Z) \right\},$$

where $c(x, w, z)$ is a cost function and Z a cost shifter.

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

- Step 3: The final outcome is given by $Y = m_0(X, q(W, \varepsilon))$.

In this setting, the agent chooses X by maximizing expected outcome, minus the cost associated with choosing X given her information set. At the time of the decision on X, the agent already observes W, which in turn was used to produce the input $U = q(W, \varepsilon)$. The variables in W can be endogenous in the sense of being dependent on ε (as is likely to be the case, since W and ε are inputs used to produce $q(W, \varepsilon)$). In addition, the agent also observes a cost shifter Z and a vector of shocks ε_x(proxies for ε observed by the agent, but not by the econometrician). Also, it is assumed that production is increasing in U (e.g., ability). What is special about this setup is that (i) production is monotonic in U; (ii) there are some observed factors W and some unobserved factors ε which only enter the production function as components of one of its inputs, q; and (iii) there exists another observable variable Z which influences the choice of X (a cost shifter), but is excluded from the production of Y, in the sense that it is neither a direct input in the production of Y, nor a direct or indirect input in the production of q ($Z \perp \varepsilon | W$). As we will show below, properties (i) to (iii) are fundamental to our identification strategy. The fact that prior inputs W do not enter final production in arbitrary ways, because of the technological constraints embodied in q, justifies our structural separability assumption between X and W, and imply the type of exclusion restrictions we exploit. In (iii), we further require that first stages are heterogeneous. Informally speaking, the cross derivative of $s(w, z, \varepsilon_x)$ in w and z is nonzero.

We now investigate identification in the nonseparable model. We show that a similar identification strategy as used for separable models allows for nonparametric identification in this more general setting. Suppose that Assumption 1 holds. Let $g_0 = m_0^{-1}$ denote the inverse of m_0 with respect to the u argument, so that $g_0(Y, X) = U$ a.s., then the exogeneity condition is

$$\mathbb{C}(g_0(Y, X), Z | W) = 0 \; a.s. \tag{31}$$

We note that these restrictions are valid for a general instrument Z, not necessarily binary. The exogeneity condition we use is generally weaker than the traditional conditional independence assumption considered in the literature

$$Z \perp U | W, \tag{32}$$

see, for example, Matzkin (2013) for discussion.

Let $\mathcal{G}$ be a subset of $L_2(Y, X)$, whose elements g satisfy the normalization restrictions $g(\bar{y}_1, \bar{x}_1) = 0$ and $g(\bar{y}_2, \bar{x}_2) = 1$, for some $\bar{x}_j \in \mathcal{S}_X, \bar{y}_j \in \mathcal{S}_Y, j = 1, 2$. Alternative normalizations can be entertained, as shown below.

Definition 4.1. *We say* (Y, X, Z) *given* W *is* $\mathcal{G}$*-covariance complete if for each* $g \in \mathcal{G} - \{g_0\}$,

$$\mathbb{C}(g(Y, X), Z | W) = 0 \Longrightarrow g = 0 \; a.s.$$

The proof of the next theorem is the same as that of Theorem 3.3 and therefore is omitted.

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

THEOREM 4.1. *Let Assumption 1 hold. Then, g_0 is point-identified in $\mathcal{G}$ if (Y,X,Z) given W is $\mathcal{G}$-covariance complete.*

Remark 4.1. Theorem 4.1 provides sufficient conditions for identification in the nonseparable case. These conditions do not exploit that $g_0(y,x)$ is monotonic in y, for each $x \in \mathcal{S}_X$. Thus, g_0 might be overidentified under the conditions of Theorem 4.1.

We extend Example 2 to the nonseparable case. For simplicity, we consider the binary IV case, as other cases can be treated similarly.

Example 8. (Example 2, cont.) Suppose $Z \in \{0,1\}$ is a binary instrument. To simplify notation denote $V = (Y,X)$ and its support by $\mathcal{S}_V := \{v_1, \ldots, v_q\}$, and let $\mathcal{S}_W := \{w_1, \ldots, w_l\}$ denote the support of W, with $q < \infty$ and $l < \infty$. Without loss of generality, we assume the normalization restrictions are $g(v_1) = 0$ and $g(v_2) = 1$. Let q_y and q_x denote the cardinality of the support of Y and X, respectively, so $q = q_y q_x$. Covariance completeness is equivalent to the rank condition

$$rank\,(\mathbf{B}) = q,$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix},$$

$\mathbf{B}_1$ is a $l \times q$ matrix with $ij-th$ element

$$b_{1ij} = \mathbb{P}[V = v_j | Z = 1, W = w_i] - \mathbb{P}[V = v_j | Z = 0, W = w_i],$$

and $\mathbf{B}_2$ is the $2 \times q$ matrix

$$\mathbf{B}_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{bmatrix}.$$

Intuitively speaking, in the nonseparable case, covariance completeness means enough heterogeneity in both reduced forms (for X and for Y), and not just on X as in the separable case (see the definition of $\mathbf{B}_1$, which involves $V = (Y,X)$). The order condition is $l \geq q - 2$.

In general, necessary conditions for covariance completeness in the nonparametric nonseparable case are that (Y,X) and W have the same level of complexity. For example, if Y is continuous but X is discrete, then W needs to be continuous, which is a different requirement than the one in separable models, where the support of Y did not play a role.

The estimation for the nonseparable case proceeds as for the separable case after suitable normalizations are imposed. The conditional exogeneity restriction leads to the moment restrictions

$$\mathbb{E}[\{g_0(Y,X) - h_0(W)\}\tilde{Z}|W] = 0 \text{ a.s.},$$

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

where now $h_0(W) = \mathbb{E}[g_0(Y,X)|W]$. This is the basis for parametric, semiparametric, or nonparametric estimators of g_0 using the general sieve estimation results of Ai and Chen (2003) and Chen and Pouzo (2012, 2015). As before, we set the intercept of g_0 to zero as location normalization. So, in the parametric case $g_0(V) = \beta'\mathbf{V}$ and $h_0(W) = \tau_0 + \gamma'\mathbf{W}$, and where now $\mathbf{V}$ and $\mathbf{W}$ are possibly large dimensional vectors of transformations of $V = (Y,X)$ and W, respectively. In the nonseparable case, we also need a scale normalization. If V_1 is one univariate component of $\mathbf{V}$ and we set to one its associated coefficient, then a TSLS estimator for the remained components of β can be computed similarly to that of the separable case: treating V_1 as the dependent variable, $-V_{-1}$ as endogenous variables (where $\mathbf{V} = (V_1, V_{-1})$) and $\mathbf{W}$ as controls. The extension to the semiparametric and nonparametric settings follows *mutatis mutandis* as in the separable case. The shape restriction provides additional identification power, and can be potentially tested under a standard TSLS rank condition. We leave the development of such tests to future research.

5. MONTE CARLO SIMULATIONS

This section investigates the finite sample performance of our TSLS estimator proposed in Section 3.3. We begin with a linear model with two basic endogenous variables and one binary instrument:

$$
\begin{aligned}
Y &= \alpha_0 + \beta_{01}X_1 + \beta_{02}X_2 + U,\\
X_1 &= \alpha_{01} + \alpha_{11}Z + \alpha_{21}W + \alpha_{31}Z\cdot W + U_1,\\
X_2 &= \alpha_{02} + \alpha_{12}Z + \alpha_{22}W + \alpha_{32}Z\cdot W + U_2,\\
U &= \gamma_0 W + u,\\
W &= \gamma_z Z + v,\\
v &= \alpha_w u + U_w,
\end{aligned}
$$

where (u, U_w, U_1, U_2) are independent standard normal random variables independent of Z, which is distributed as Bernoulli random variables with probability $p = 0.5$. The classical order condition of standard IV does not hold in this model, and hence, classical IV is unable to identify the marginal effects β_{01} and β_{02}. For identification of the marginal effects, our method requires that the conditional exogeneity assumption holds. Since Z is binary, (11) is equivalent to (10) (cf. Proposition 3.1). Note that in this data generating process (DGP1, henceforth) some simple calculations yield

$$
\begin{aligned}
\mathbb{E}[U|W,Z] &= \gamma_0 W + \mathbb{E}[u|W,Z]\\
&= \gamma_0 W + \frac{\alpha_w(W - \gamma_z Z)}{1+\alpha_w^2}.
\end{aligned}
$$

Thus, the conditional exogeneity holds provided

Case 1: $\alpha_w = 0$ and/or Case 2: $\gamma_z = 0$.

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

Table 2. Mean squared error for DGP1

α_{11}	α_{31}	α_{12}	α_{32}	n	Case 1 (β_1)	Case 2 (β_1)	Case 1 (β_2)	Case 2 (β_2)
1	0	0	1	100	0.1622	0.0251	0.0981	0.0122
				300	0.0217	0.0070	0.0149	0.0035
				500	0.0124	0.0041	0.0085	0.0020
				1,000	0.0062	0.0020	0.0041	0.0010
0	1	1	0	100	0.5941	0.0121	0.8514	0.0250
				300	0.0147	0.0035	0.0221	0.0071
				500	0.0085	0.0020	0.0127	0.0041
				1,000	0.0042	0.0010	0.0061	0.0020
1.5	1	1	1.5	100	262.2022	0.3708	246.6096	0.2776
				300	0.7721	0.0136	0.6172	0.0105
				500	0.0392	0.0079	0.0351	0.0061
				1,000	0.0164	0.0036	0.0147	0.0028

Note: 10,000 Monte Carlo simulations.

The relevance condition will hold when (7) is satisfied, that is, $\alpha_{11}\alpha_{32} \neq \alpha_{31}\alpha_{12}$. The parameters in the structural equation are set at $\alpha_0 = 0, \beta_{01} = 1, \beta_{02} = 2,$ and $\gamma_0 = 1$ (so W is endogenous in the sense of this paper). Table 2 provides the Mean Squared Error (MSE) based on 10,000 Monte Carlo simulations, sample sizes $n = 100, 300,$ 500, and $1,000$, and several values for $(\alpha_{11}, \alpha_{31}, \alpha_{12}, \alpha_{32})$ and (α_w, γ_z). There are three variables in the structural equation

$$Y = \alpha_0 + \beta_{01}X_1 + \beta_{02}X_2 + \gamma W + \varepsilon,$$

and three in each reduced form equation, and therefore, the TSLS estimator is an IV estimator that treats Z and the interaction term $Z \cdot W$ as instruments. We note that Table 2 does not offer a comparison with existing methods such as IV because they are not applicable (e.g., classical IV's order condition fails).

We consider two cases, Case 1: $\alpha_w = 0$ and $\gamma_z = 1$ and Case 2: $\alpha_w = 1$ and $\gamma_z = 0$, and for each case, three subcases corresponding to $(\alpha_{11} = 1, \alpha_{31} = 0, \alpha_{12} = 0, \alpha_{32} = 1)$, $(\alpha_{11} = 0, \alpha_{31} = 1, \alpha_{12} = 1, \alpha_{32} = 0)$, and $(\alpha_{11} = 1.5, \alpha_{31} = 1, \alpha_{12} = 1, \alpha_{32} = 1.5)$. We refer to the first two subcases as strongly identified and the last one as moderately identified (the rank condition is strongly satisfied in the first two subcases). In these simulations, we focus on MSE. Unreported simulations suggest a satisfactory bias performance uniformly over all parameter values. Thus, the observed variations in MSE come mainly from differences in the estimators' variances. Several observations arise from the reported results. First, Case 1 leads to larger MSEs than Case 2. Our interpretation for this is that covariance completeness holds more strongly in Case 2 than in Case 1. To see this, note that simple calculations yield

$$\mathbb{V}(W) = \gamma_z^2 p(1-p) + \alpha_w^2 + 1,$$

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

and the propensity score

$$p_0(w) = \frac{p\phi\left(\frac{w-\gamma_z}{\sqrt{(1+\alpha_w^2)}}\right)}{p\phi\left(\frac{w-\gamma_z}{\sqrt{(1+\alpha_w^2)}}\right) + (1-p)\phi\left(\frac{w}{\sqrt{(1+\alpha_w^2)}}\right)}.$$

In Case 2, both $\mathbb{V}(W)$ and $\mathbb{V}[Z|W]$ are larger, since $p_0(w) \equiv p = 1/2$. The rank condition (25) holds more strongly in Case 2 than in Case 1, while the rank condition (24) remains the same in each subcase. Thus, covariance completeness is stronger in Case 2. These results also show that endogeneity of W, in the sense of W being dependent with U or u, does not affect negatively the MSE.

For the first two subcases ($\alpha_{11} = 1, \alpha_{31} = 0, \alpha_{12} = 0, \alpha_{32} = 1$) and ($\alpha_{11} = 0, \alpha_{31} = 1, \alpha_{12} = 1, \alpha_{32} = 0$), the sample variance of the estimators is already small for small sample sizes as 100, and it decreases to zero with the sample size, in accordance with the consistency of the estimator. For the parameter values ($\alpha_{11} = 1.5, \alpha_{31} = 1, \alpha_{12} = 1, \alpha_{32} = 1.5$), identification is weaker, and larger sample sizes than $n = 100$ are required for a good performance, particularly for Case 1. In all cases, the MSE decreases with the sample size. Overall, these results provide supporting evidence of the robustness of our identification strategy to the endogeneity of W and that the finite sample performance depends on the strength of the covariance completeness condition in terms of both rank conditions, (24) and (25). Unreported results confirm that if the conditional exogeneity condition does not hold then our estimators may be biased and inconsistent. Consider, for example, the case $\alpha_w = 1$ and $\gamma_z = 1$, and the first configuration ($\alpha_{11} = 1, \alpha_{31} = 0, \alpha_{12} = 0, \alpha_{32} = 1$). Then, it can be shown that β_{01} is not identified, since (with the generic notation $C_\xi(W) = \mathbb{C}(\xi, Z|W)$)

$$C_Y(W) = \beta_{01} C_{X_1}(W) + \beta_{02} C_{X_2}(W) + C_\varepsilon(W)$$

and $C_\varepsilon(W) = \rho\gamma_z C_{X_1}(W)$, so there is multicollinearity in the last regression. However, our estimator for β_{02} is consistent even in this case.

In the second set of simulations, we consider a semiparametric version of our estimator where h_0 is left unspecified. We generate data as above, but replace the equation for U by $U = \gamma_0 \sin W + u$, which corresponds to the specification

$$h_0(W) = \rho W + \gamma \sin W.$$

We then use our estimator with $\mathbf{W} = (q_2(w), \ldots, q_{J_n}(w))'$, for power series $q_j(w) = w^{j-1}$ and several values of $d_w = J_n - 1$. The parameters are set to $\alpha_w = 0, \gamma_z = 1, \alpha_{11} = 1, \alpha_{31} = 0, \alpha_{12} = 0, \alpha_{32} = 1$ (corresponding to Case 1 with the first subcase above), and other parameters remaining the same: $\alpha_0 = 0, \beta_{01} = 1, \beta_{02} = 2$ and $\gamma_0 = 1$. We consider three choices for the number of elements in the power series basis: "low" ($d_w = 1$), "moderate" ($d_w = 3$), and "high" ($d_w = 6$). Table 3 provides the average bias and MSE based on 10,000 Monte Carlo simulations for these parameter configurations. The misspecification bias leads to a large estimation bias

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

TABLE 3. Semiparametric results: Case 1

d_w	n	Bias (β_1)	MSE (β_1)	Bias (β_2)	MSE (β_2)
1	100	0.2472	12.3415	−0.3222	3.6223
	300	0.1884	0.0654	−0.2856	0.1049
	500	0.1883	0.0520	−0.2832	0.0938
	1,000	0.1850	0.0424	−0.2805	0.0850
3	100	0.0051	0.0921	−0.0159	0.0932
	300	0.0086	0.0249	−0.0169	0.0244
	500	0.0129	0.0141	−0.0215	0.0139
	1,000	0.0125	0.0072	−0.0203	0.0070
6	100	−0.0053	0.0609	0.0066	0.0641
	300	0.0016	0.0215	0.0001	0.0205
	500	0.0002	0.0131	0.0001	0.0126
	1,000	0.0000	0.0065	−0.0004	0.0061

Note: 10,000 Monte Carlo simulations.

for $d_w = 1$ but a small bias already for $d_w = 3$. Including more controls leads to a reduction in variance, as can be seen from the large reduction in variance for $n = 100$. For larger sample sizes, the MSE is relatively stable for $d_w = 3$ and $d_w = 6$. Unreported results with other subcases lead to qualitatively the same conclusions.

Table 4 provides the average bias and MSE based on 10,000 Monte Carlo simulations for Case 2: $\alpha_w = 1$ and $\gamma_z = 0$. Now the misspecification bias does not lead to estimation bias. This can be explained by the independence of the instrument Z and the control W, which makes the estimation robust to misspecification bias in h_0.

In sum, these simulations provide finite sample evidence of a satisfactory performance of the TSLS estimator and its robustness to the endogeneity of the controls and the misspecification bias in h_0 when it is implemented semiparametrically. It has been shown that, for the finite sample performance, covariance completeness in the form of both rank conditions (24) and (25) are important.

6. AN APPLICATION TO THE ESTIMATION OF THE EFFECTS OF AIR POLLUTION ON HOUSE PRICES

We apply our identification strategy to the problem of estimating the effects of pollution on house prices, as in Chay and Greenstone (2005). The concern with endogeneity in this problem is warranted, since counties may differ from each other in many ways which may not be accounted by their observable characteristics and amenities. Chay and Greenstone base their identification strategy on an IV approach, which takes advantage of the quasi-experiment generated by the Clean Air Act around the time it was first implemented.

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

Table 4. Semiparametric results: Case 2

d_w	n	Bias (β_1)	MSE (β_1)	Bias (β_2)	MSE (β_2)
1	100	−0.0007	0.0359	−0.0027	0.0311
	300	0.0005	0.0101	−0.0006	0.0093
	500	0.0014	0.0060	0.0011	0.0055
	1,000	0.0007	0.0029	0.0001	0.0028
3	100	0.0005	0.0235	0.0008	0.0151
	300	−0.0001	0.0076	0.0008	0.0047
	500	0.0001	0.0045	−0.0003	0.0028
	1,000	−0.0000	0.0023	0.0005	0.0014
6	100	0.0015	0.0206	−0.0018	0.0128
	300	−0.0014	0.0066	0.0005	0.0037
	500	−0.0004	0.0041	−0.0004	0.0021
	1,000	0.0006	0.0020	0.0002	0.0010

Note: 10,000 Monte Carlo simulations.

Let Y denote the change between 1970 and 1980 in the logs of the county's median property values, X is the change between 1970 and 1980 in the geometric mean total suspended particulates (TSPs) across all monitors in the county and Z is the county's attainment status in 1975 according to the Clean Air Act. Since Chay and Greenstone use a rich set of controls (see p. 411 in that paper), we separate them into two vectors Q and W as explained in Section 3.2.1.

Our first concern when choosing W is the implications it has for the model in terms of the specific functional shapes as well as structural separability. To begin, we follow the guidance set out in Example 7. The county's changes between 1970 and 1980 in the percent spending in highways, health and education do not affect the valuation for air quality directly, though they can be influential through affecting U (e.g., when they cause a change in the composition of the county's population). Thus, the three variables above are prime candidates to serve as W (see considerations about the choice of W and robustness checks below). We began by considering 4 models with the structure $Y = \beta X + Q'\pi + U$, with $\mathbb{E}[U|Z, Q, W] = \mathbb{E}[U|W]$. The first three use a scalar W (each of the three variables above), and in the fourth W is a vector of all three variables at the same time (see Table 5). These models have a similar structure to that of Chay and Greenstone (2005), which allows us to do direct comparisons.

The estimation is done using the same data set as in Chay and Greenstone (2005), and Q is the exact same control list used in that paper. The first column in Table 5 shows the results of a standard IV estimation of β, which is what is done in Chay and Greenstone (2005). The replicated results are, not surprisingly, identical to that paper. Columns A to D show the results of our estimation approach using different variables as the separable classification control W. Row

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

TABLE 5. Estimation results—Linear case, binary IV

	IV	A	B	C	D
(i)	−.347	−.340	−.327	−.317	−.327
β, no controls	(.140)	(.138)	(.136)	(.134)	(.135)
(ii)	−.203	−.208	−.202	−.203	−.208
β, with controls	(.093)	(.094)	(.093)	(.093)	(.093)
(iii)		−.226	−.220	−.194	
β in nonseparable model		(.096)	(.098)	(.110)	
(iv)		−1.432	.854	.174	
β_1 (separability test)		(1.873)	(1.227)	(1.174)	

Note: Columns A to D use our approach with W equal to the change from 1970 to 1980 in the % spending on highways (A), health (B), and education (C). In Column (D), W, is the vector of all three variables. Specification (i) has no exogenous controls Q. Specification (ii) uses as exogenous controls Q the exact same specification as in Chay and Greenstone (2005) excluding the controls which we are using as W. Specification (iii) assumes a nonseparable model as described in Example 5. Row (iv) shows the results of the separability test described in Section 3.2.1.

(i) uses a specification without exogenous control variables ($\pi = 0$). Although in specification (i) our results are of similar magnitude to the standard IV, they are slightly smaller and vary depending on the chosen W. We believe that this happens because when controls Q are not used, the classic IV operates under the assumption that $\mathbb{E}[U|Z]$ is constant, while our estimator operates under the assumption that $\mathbb{E}[U|Z,W] = \mathbb{E}[U|W]$. That said, Chay and Greenstone never suppose that their IV is valid, but only that it is valid conditional on controls. Row (ii) shows the results conditional on controls. There the identifying assumption of the classic IV approach is that $\mathbb{E}[U|Z,Q,W]$ is constant, while our identifying assumption is that $\mathbb{E}[U|Z,Q,W] = \mathbb{E}[U|W]$. Nevertheless, our results generally confirm the estimates found by Chay and Greenstone.

As a robustness check of the separability of the three variables above we followed Example 5 to estimate the model $Y = \beta(W)X + Q'\pi + U$, with $\beta(W) = \beta + \beta_1' W$, under our assumption $\mathbb{E}[U|Z,Q,W] = \mathbb{E}[U|W]$. The results are very similar and can be seen in Row (iii). Also, as explained in Section 3.2.1, we can test the separability of W and X by testing whether $\beta_1 = 0$. Row (iv) shows the estimates $\hat{\beta}_1$, which are all insignificant, and thus we are confident that our choice of W satisfies the structural separability requirement of our method.

To further test the separability of W assumption, we also ran the same regressions in Row (ii) using each of the other controls in the model as W instead. The results are extremely similar. The five controls which yielded the most different results are the number of houses built between 1970 and 1980, the rate of vacancies in 1980, change in income per-capita, the change in government revenue per-capita and the change in the fraction of the population with at least a college degree, but even in these cases the differences between our estimates and −.203 was always

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

less than .1. In comparison, Chay and Greenston's Regression Discontinuity Design (RDD) estimate is $-.275$, which they consider to be a confirmation of their results. We note that the observed robustness of our method to the specification of W in this application can be theoretically justified by the over-identification of our method even in cases where classical IV just-identifies.

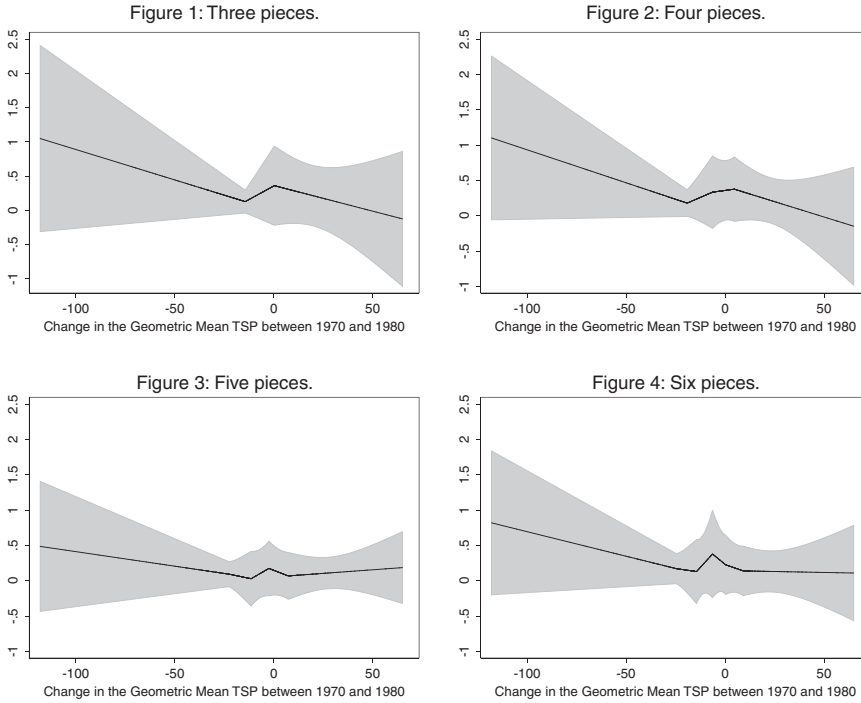
The comparative advantages of our method are better showcased in the nonlinear case, where the classical IV methods cannot identify marginal effects with a single IV . Given our test confirming the separability of W, we restrict ourselves to the model $Y = g(X) + Q'\pi + U$, with $\mathbb{E}[U|Z, Q, W] = \mathbb{E}[U|W]$, where g is a connected piecewise linear function, and thus X can have richer marginal effects on Y. This model allows us to compare our results directly with those of Chay and Greenstone. Figure 1 has three linear pieces, which connect at the terciles of the distribution of X. Hence, we can write $g(x) = a_1\psi_1^3(x) + a_2\psi_2^3(x) + a_3\psi_3^3(x)$, where the $\psi_j^3(x)$ are the elements of a B-spline basis of degree 1 and smoothness 0 with knots at the terciles just described. In practice, this is the same as if we had three endogenous variables $\psi_1^3(X)$, $\psi_2^3(X)$ and $\psi_3^3(X)$. For W, we used the three variables in Column D in Table 5 (call them *High*, *HealZh* and *Educ*) expanded into the elements of the B-spline basis. So, with some abuse of notation $W = (\psi_1^3(High), \psi_2^3(High), \psi_3^3(High), \psi_1^3(HealZh), \psi_2^3(HealZh), \psi_3^3(HealZh), \psi_1^3(Educ), \psi_2^3(Educ), \psi_3^3(Educ))'$. Figures 2–4 are obtained analogously.

In Figure 1, the standard errors are calculated as the standard error of the predicted $\hat{g}(x) = \hat{a}_1\psi_1^3(x) + \hat{a}_2\psi_2^3(x) + \hat{a}_3\psi_3^3(x)$, so $SE(\widehat{g(x)}) = (\psi_1^3(x), \psi_2^3(x), \psi_3^3(x))'$ $\Omega(\psi_1^3(x), \psi_2^3(x), \psi_3^3(x))$, where Ω is the estimated covariance matrix of $(\hat{a}_1, \hat{a}_2, \hat{a}_3)'$. In a prepackaged software, the standard errors can be obtained directly as the standard errors of the predicted g. Standard errors in Figures 2–4 are obtained analogously. Note that uniform confidence bands can be obtained from the pointwise confidence interval by simply extending the widest radius among the point estimate standard errors to the entire curve.

Figures 1–4 are representative of the patterns we found when we tried other strategies. For example, we also used as W each of the elements used in Columns A to C in Table 5 separately (i.e., for g with three pieces we specified $W = \left(\psi_1^3(High), \psi_2^3(High), \psi_3^3(High)\right)'$, and we also expanded the elements of W in other ways, for example, into two piece B-splines ($W = (\psi_1^2(High), \psi_2^2(High), \psi_1^2(HealZh), \psi_2^2(HealZh), \psi_1^2(Educ), \psi_2^2(Educ))'$), four piece B-splines, etc., all with very similar results. The standard errors get substantially larger as we increase the number of pieces in g, but they are not affected (and seem in fact to decrease) as we increase the number of elements in W.

Our application on the effect of air pollution on house prices confirms the findings of Chay and Greenstone (2005) when a constant marginal effect model is considered, and uncovers a potentially interesting heterogeneity in the effect of pollution on house prices when richer marginal effects are entertained. The impact of air quality on house prices tends to be larger for counties that substantially change their pollution levels as a result of the Clean Air Act relative to those that experience only a minor decrease in pollution during the period. A possible

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press



Note: Nonlinear case—IV approach. Curves are the results of our approach with exogenous controls, and Z as in Column D in Table 5. The domain of each piece is determined by the quantiles. For instance, Figure 2's pieces connect at the 25th, 50th, and 75th quantiles of the change in the geometric Mean TSP between 1970 and 1980.

interpretation is that only large reductions in pollution trigger enough differential sorting of residents to generate the positive effects on price. However, large standard errors prevent us from rejecting the linear specification.

7. CONCLUSIONS

In this paper, we have proposed a strategy to identify marginal effects of endogenous variables with richer support than instruments, in the presence of other possible endogenous controls. The strategy hinges on the heterogeneity of the "first stages" and a structural separability of some controls, and it can be extended to nonseparable models with unobserved marginal effects. It can be applied to parametric, semiparametric, and nonparametric settings. The identification strategy can be implemented with a simple TSLS estimator that treats the controls as if they were exogenous, and runs a first stage with interactions between the IV and the controls. Thus, our identification strategy can be readily implemented with off-the-shelf econometric software. Monte Carlo simulations show that this TSLS

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

estimator performs well in practice, and it is robust to endogeneity of controls and misspecification of the nuisance parameters.

There are several extensions of our methods that deserve further investigation. First, as the first version of this paper shows, the proposed methods can be extended to the RDD setting. In that setting, we have devised a practically convenient semiparametric estimator that uses a varying coefficients specification of the first stages. Identification and semiparametric estimation in the RDD setting will be investigated in a companion paper.

References

Ai, C. & X. Chen (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71(6), 1795–1843.

Almond, D. & J. Currie (2011) Killing me softly: The fetal origins hypothesis. *The Journal of Economic Perspectives* 25(3), 153–172.

Altonji, J.G. & R.L. Matzkin (2005) Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* 73, 1053–1102.

Andrews, D. (2017) Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics* 199, 213–220.

Ben-Moshe, D., X. D'Haultfoeuille & A. Lewbel (2017) Identification of additive and polynomial models of mismeasured regressors without instruments. *Journal of Econometrics* 200, 207–222.

Blundell, R., X. Chen & D. Kristensen (2007) Semi-nonparametric IV estimation of shape-invariant Engel curves *Econometrica* 75, 1613–1670.

Caetano, C., J.C. Escanciano & A. Bergman (2019) MMEIV: Stata module to perform Multiple Marginal Effects IV Estimation. Statistical Software Components S458674, Boston College Department of Economics.

Card, D. (2001) Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* 69, 1127–1160.

Chay, K. & M. Greenstone (2005) Does air quality matter? Evidence from the housing market. *Journal of Political Economy* 113(2), 376–424.

Chen, X. & T. Christensen (2018) Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics* 9(1), 39–84.

Chen, X. & D. Pouzo (2012) Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* 80(1), 277–321.

Chen, X. & D. Pouzo (2015) Sieve Wald and QLR inference on semi/nonparametric conditional moment models. *Econometrica* 83(3), 1013–1079.

Chernozhukov, V. & C. Hansen (2005) An IV model of quantile treatment effects. *Econometrica* 73, 245–261.

Chesher, A. (2003) Identification in nonseparable models. *Econometrica* 71, 1405–1441.

D'Haultfoeuille, X. (2011) On the completeness condition in nonparametric instrumental problems. *Econometric Theory* 1, 1–12.

D'Haultfoeuille, X. & P. Fevrier (2015) Identification of nonseparable models with endogeneity and discrete instruments. *Econometrica* 83, 1199–1210.

D'Haultfoeuille, X., S. Holderlein & Y. Sasaki (2013) Nonlinear Difference-in-Differences in Repeated Cross Sections with Continuous Treatments, Boston College Working Paper wp839.

Darolles, S., Y. Fan, J.P. Florens & E. Renault (2011) Nonparametric instrumental regression. *Econometrica* 79(5), 1541–1565.

Dunker, F., J.-P. Florens, T. Hohage, J. Johannes & E. Mammen (2014) Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *Journal of Econometrics* 178, 444–455.

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

Florens, J.P., J.J. Heckman, C. Meghir & E. Vytlacil (2008) Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effect. *Econometrica* 76, 1191–1207.

Frolich, M. (2008) Parametric and nonparametric regression in the presence of endogenous control variables. *International Statistical Review* 76, 214–227.

Horowitz, J. (2011) Applied nonparametric instrumental variables estimation. *Econometrica* 79(2), 347–394.

Horowitz, J. (2014) Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics* 180, 158–173.

Huang, L., U. Khalil & N. Yildiz (2019) Identification and estimation of a triangular model with multiple endogenous variables and insufficiently many instrumental variables. *Journal of Econometrics* 208, 346–366.

Kim, K.I. & S. Song (2017) Estimation of Semiparametric Models with Mismeasured Endogenous Regressors Using Control Variables. Working Paper.

Lumley, J., C. Chamberlain, T. Dowswell, S. Oliver, L. Oakley & L. Watson (2009) Interventions for promoting smoking cessation during pregnancy (Cochrane Review). *The Cochrane Library* 8, 3.

Masten, M. & A. Torgovitsky (2016) Identification of instrumental variable correlated random coefficients models. *The Review of Economics and Statistics* 98(5), 1001–1005.

Matzkin, R.L. (2013) Nonparametric identification in structural economic models. *Annual Review of Economics* 5, 457.

Newey, W.K. & J. Powell (2003) Instrumental variables estimation for nonparametric models. *Econometrica* 71, 1565–1578.

Stock, J.H. & M.W. Watson (2011) *Introduction to Econometrics*, 3rd ed. Addison Wesley.

Torgovitsky, A. (2015) Identification of nonseparable models using instruments with small support. *Econometrica* 83, 1185–1197.

Wahba, G. (1990) *Spline Models for Observational Data*. SIAM.

APPENDIX

A.1. Proofs of the Main Results

Proof of Proposition 3.1: Trivially,

$$\mathbb{C}(U,Z|W) = \mathbb{E}[\{U - \mathbb{E}[U|W]\}Z|W],$$

which implies (10), is equivalent to the conditional moment restriction

$$\mathbb{E}[\{Y - g_0(X) - h_0(W)\}\tilde{Z}|W] = 0 \text{ a.s.}, \tag{A.1}$$

where $\tilde{Z} = (1,Z)$ and $h_0(W) = \mathbb{E}[U|W]$. Alternatively, (11) is equivalent to

$$\mathbb{E}[Y - g_0(X) - h_0(W)|W,Z] = 0 \text{ a.s.} \tag{A.2}$$

By the law of iterated expectations the last condition implies (A.1). For (ii), note that (A.2) is equivalent to

$$\mathbb{E}[(Y - g_0(X) - h_0(W))\, r(W,Z)] = 0 \tag{A.3}$$

for all measurable functions r such as the previous moment is well-defined. In the binary case, such r can be written as

$$\begin{aligned} r(W,Z) &= Zr(W,1) + (1-Z)r(W,0) \\ &= r(W,0) + Z(r(W,1) - r(W,0)). \end{aligned}$$

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press

Substituting this r into the moment function yields

$$\mathbb{E}[(Y-g_0(X)-h_0(W))\,r(W,Z)] = \mathbb{E}[(Y-g_0(X)-h_0(W))\,r(W,0)] + \mathbb{E}[(Y-g_0(X)-h_0(W))\,Z(r(W,1)-r(W,0))].$$

Both of these moments are zero under (A.1), which proves (ii). ■

Proof of Theorem 3.2: Note that $A\widetilde{g} = \Delta_Y$, with $\widetilde{g}(X) = g_0(X) - \mathbb{E}[g_0(X)]$ and A is invertible on the orthocomplement of $\mathcal{N}(A)$, which by Assumption 1 is given by $\mathcal{N}^{\perp} = \{\lambda \in \mathcal{G} : \mathbb{E}[\lambda(X)] = 0\}$. Thus, since $\widetilde{g} \in \mathcal{N}^{\perp}$, it holds that $\widetilde{g} = A^{-1}\Delta_Y$ and therefore $\widetilde{g}$ is identified. ■

Proof of Theorem 3.3: Assumption 1(ii) gives

$$\mathbb{C}(Y,Z|W) = \mathbb{C}(g_0(X),Z|W). \tag{A.4}$$

Suppose (X,Z) given W is $\mathcal{G}$-covariance complete and let $g_1 \in \mathcal{G}$ such that

$$\mathbb{C}(Y,Z|W) = \mathbb{C}(g_1(X),Z|W) \text{ a.s.}$$

Thus, $\mathbb{C}(g_1(X)-g_0(X),Z|W) = 0$ a.s. Then, covariance completeness implies $g_0 = g_1$ a.s. This proves identification. The reciprocal also holds. Suppose that (X,Z) given W is not $\mathcal{G}$-covariance complete, then we can find $g_1 \in \mathcal{G} - \{g_0\}, g_1 \neq 0$, such that $\mathbb{C}(g_1(X),Z|W) = 0$. Then, identification of g_0 in $\mathcal{G}$ fails, because $g_2 \equiv g_0 + g_1 \in \mathcal{G}$ and g_2 satisfies (A.4). ■

Proof of Corollary 3.1: For $Ag := \mathbb{E}[g(X)|W,Z=1] - \mathbb{E}[g(X)|W,Z=0]$, arguing as in Proposition 3.1, we can write

$$\begin{aligned}\mathcal{N}(A) &= \{g : \mathbb{E}[\{g(X)-\mathbb{E}[g(X)|W]\}|W,Z] = 0\}\\ &= \{g : \mathbb{E}[\{g(X)-\mathbb{E}[g(X)|W]\}\tilde{Z}|W] = 0\}.\end{aligned}$$

Assumption 2 states $\mathcal{N}(A)$ is made only of constants, and the location normalization imposes the constant must be zero. This is covariance completeness. In the other direction, suppose there exists a non-constant g in $\mathcal{N}(A)$. If $g(\bar{x}) = 0$, then covariance completeness fails. If $g(\bar{x}) \neq 0$, we can construct the function $g_1(x) = g(x) - g(\bar{x})$, which satisfies $g_1 \in \mathcal{N}(A)$ and $g_1(\bar{x}) = 0$ (note that since g is non-constant g_1 is non-zero). Again, this shows covariance completeness fails. ■

Proof of Proposition 3.4: Write, by the law of iterated expectations

$$\begin{aligned}\mathbb{C}(g(X),Z|W) &= \mathbb{E}[g(X)\{Z-\mathbb{E}[Z|W]\}|W,Z]\\ &= \mathbb{E}[g(X)k(X,W)|W,Z].\end{aligned}$$ ■

https://doi.org/10.1017/S0266466620000213 Published online by Cambridge University Press